

Filterprints: Identifying Localised Usage Anomalies in Censorship Circumvention Tools

Joss Wright, Alexander Darer, Oliver Farnan
joss.wright@oii.ox.ac.uk
alexander.darer@linacre.ox.ac.uk
oliver.farnan@balliol.ox.ac.uk

August 11, 2016

Abstract

Motivated by the desire to detect automatically instances of internet filtering and censorship, we propose an approach based on principle component analysis to identify per-country anomalous periods in the usage of censorship circumvention tools, and demonstrate the applicability of this approach with global usage statistics from the Tor Project. Our technique is presented as a tool that automatically highlights periods of anomalous usage of Tor on a per-country basis and ranks countries' level of anomalous behaviour over given time periods, but is also applicable to usage data of other services. In contrast to previous country-specific investigations, the technique presented here uses deviation from expected patterns of behaviour, calculated on a per-country basis, to identify countries whose usage of Tor have deviated from those of similarly behaving countries. We demonstrate the effectiveness of this approach against known historical filtering events as well as synthetically injected anomalous events, and evaluate the sensitivity of this technique against different classes of usage anomaly. Our results suggest that patterns of usage of circumvention tools, such as the Tor Project, act not only as direct indicators of network filtering but also as a meaningful proxy variable for related events, such as protests or political unrest, in which internet use is restricted.

1 Introduction

Nation states, and others, have increasingly begun to employ internet filtering as a means of controlling access to information, and as a tool to limit certain forms of social and political organisation. Given the central role that the internet plays in communications for a large and increasing proportion of the global population, understanding the application and development of filtering technologies, and the effects of these methods on individuals and society, is of great importance. Whilst analyses of known filtering infrastructures provide useful data for identifying tools, techniques, and limitations of these entities, discovering internet filtering behaviour in less-studied regions is of great importance.

Much existing research into internet filtering has focused either on observing practices of states already known to engage in filtering, or in the development of censorship circumvention tools. Whilst multilateral studies of censorship have been conducted, most notably the seminal work of Deibert et al. [7], these approaches have typically amalgamated manual country-specific investigations. In the case of [7], countries were hand-ranked according to a number of broad criteria for internet freedom, based on network measurements as well as media reporting and expert interviews.

A core problem in research into network filtering is the discovery of filtering events and their targets. To

date, most technical analyses of filtering have focused on known filtering countries, or on word of mouth or media reports regarding new events. The work presented here is intended to provide a means to alert researchers and activists to potential developing events that may otherwise have been missed, by focusing on patterns of circumvention tool usage around the world.

We highlight that blind analysis of patterns of circumvention tool usage cannot themselves prove that a state is engaging in network filtering. A variety of different events affect the usage of circumvention tools: political unrest, media reports, elections, natural disasters, and many others could cause users to make use of such tools for a variety of reasons. One major form of event, however, that affects the usage of circumvention tools will be the imposition, or relaxation, of internet filtering and the corresponding rise or fall in people who choose to adopt such a tool.

The work presented here is motivated by a desire to identify global patterns of internet filtering¹ through automated approaches based on network data, and to link these events to their broader social and political context. In the current work we focus on the former whilst considering the means to achieve the latter.

1.1 Contributions

This work makes both theoretical contributions to anomaly detection, specifically in circumvention tool usage; and a practical contributions in the form of an implemented tool for detecting anomalous events in Tor usage data. Specifically, we make the following practical contributions:

- A tool to detect and highlight anomalous periods in per-country usage of the Tor network, making use of published metrics;
- a ranking of the most anomalous countries, in terms of usage, over a given time period;

¹The term *censorship* is commonly used in the field to refer to manipulation of network traffic for social or political purposes. To avoid making normative judgements on the nature of particular events, we prefer the more neutral terms *filtering*, *blocking*, or *manipulation* when possible.

Our practical contribution is built on our key theoretical contribution:

- An application of principal component analysis to detecting and quantifying anomalous periods of per-country usage, applied over high-dimensional, non-stationary, time series data.

We validate our approach’s effectiveness through detecting both a range of test anomalies artificially injected into the data, and known reported filtering events against the Tor network.

As an addition to the above, we suggest that usage data for tools such as the Tor project, and others, can act as practical and effective proxy variables for identifying a wider class of internet-focused social and political events around the world.

1.2 Problem and Approach

When an entity, such as a state, chooses to filter or block certain types of information, the resulting patterns of traffic reflect the intervention in the form of statistical anomalies. In a global system, in which many entities may be interfering with traffic or publicising their attempts to do so, it is desirable to identify *localised* anomalies and to gain an understanding of their nature.

We seek to detect anomalous behavior in patterns of circumvention tool traffic, motivated by the desire to identify, directly or indirectly, filtering practices; as such, we focus on usage data from circumvention tools, most notably the Tor project. Despite our focus on filtering practices, the usage patterns of a tool such as Tor are also affected by a range of exogenous factors such as social and political unrest, genuine network outages, and media reporting. It is therefore hard to assert any form of causal relationship between observed anomalies and technical filtering interventions. The work presented here should therefore more correctly be identified as detecting statistical anomalies in circumvention tool usage to provide a series of *indicators* that a certain state is worthy of further investigation over a given time period.

We consider the problem of detecting events of interest from the perspective of anomalous behaviour

in traffic flows. Key to our approach is the modelling of each country’s behaviour as a function of the behaviour of other countries, with time periods judged as anomalous if they deviate from these models. More intuitively: if a country typically exhibits similar patterns of usage to some subset of other countries, it is of interest when its usage patterns start to deviate significantly from those other countries’ usage.

The intuition underlying this work is that, in the absence of interference, certain network statistics for countries are likely to fall into one of a small number of classes, and that this classification will remain relatively stable over time. If an individual country begins to deviate from its prior classification, it suggests that a key factor regulating that usage has altered; this may be a filtering intervention, such as the blocking of a major service or tool. Of course, this may also indicate some other technical, or non-technical, factor such as a network outage causing anomalous behaviour.

In our approach, we employ *principal component analysis* to construct a continually-updated model of each country’s observed usage. This model can be considered a simplified linear combination of per-country statistics that approximates the original data, with minor variances falling outside the model. By differencing the observed usage data against the approximate model, we produce a series of per-country *residuals* that correspond to localised anomalies with respect to the modelled behaviour.

The fundamental technique underlying this approach, which splits the set of *principal components* into *normal* and *anomalous* subspaces, was initially proposed by Jackson and Mudholkar [13] for application in industrial process control. It was later employed by Lakhina et al. [17] to detect overall network-wide traffic anomalies from per-link data in high-performance networks. In contrast our approach focuses not on network-wide conditions, but on per-country variations, and explicitly accounts for long-term evolution of the data.

We adapt the subspace anomaly technique to time series data by applying it over a rolling time window, and make use of robust statistics [12] to correct for long-term shifts in usage and to specify adaptively the per-country conditions that should be considered

as anomalous.

We discuss the technique of principal component analysis in §3.2, and the details of the PCA-subspace methodology in §3.2.2.

2 Existing Work

Internet filtering, and more broadly censorship, has received attention from a number of fields of study. Technical research has focused on analysing mechanisms of censorship, and the development of censorship circumvention approaches. At the same time, researchers from the social sciences have investigated the motivations of censors, and the legal, economic, and societal effects of such systems. We argue that a holistic understanding of internet filtering, and the interaction between technical capabilities and human factors, is necessary to influence its future development.

2.1 Technical Analyses

Arguably the most well-known national-level filtering system is that of China, commonly known as the Great Firewall. One of the earliest significant studies of this system was presented by Clayton et al. [4], who isolated one mechanism by which connections were interrupted if particular keywords were identified in traffic. The mechanism discovered by Clayton et al. resulted in TCP RST packets being sent from an intermediary router to both source and destination of a connection if a filtering criterion was met. The authors further demonstrated that if the two endpoints of the connection ignored the TCP RST, the connection could successfully continue.

In more recent work, it has become apparent that the Chinese approach to filtering is both complex and evolving. In two recent papers, a group of anonymous researchers have explored manipulation, or poisoning, of DNS records that pass through China [2, 3]. This work has identified DNS manipulation as one of the most prevalent forms of filtering in China. Similarly, Wright [30] demonstrated that DNS censorship was experienced differently in different regions within

China, with significant variation in the nature of the DNS poisoning seen across the country.

Crandall et al.[5] make use of *latent semantic analysis* to derive, from known terms blocked in HTTP traffic going into China, semantically related keywords that might also be blocked. These derived keywords can then be verified by the simple process of attempting to make HTTP connections into China containing the suspect words. This approach aims to produce a continually-updated list of blocked terms that could be used to maintain an understanding of those terms most offensive to the filtering authorities.

2.2 Global Studies

Perhaps the most comprehensive study to date of global filtering practices is given by Deibert et al. [7]. In this work the authors carried out a range of remote and in-country analyses over a number of years, incorporating both technical measurements and interviews with local experts. The resulting research presented a series of snapshots of individual countries, with both an overview of the social, political, and technical landscape, and censorship practices rated on a simple scale in various categories of content: political, social, conflict and security, and internet tools.

Whilst the approach of [7] is far more comprehensive in its scope than other studies, it relies on a largely manual process that would require significant ongoing resources to maintain as a continuous overview of the state of internet filtering.

More recently, the OONI project[22] has developed a platform to conduct a variety of tests for network filtering, with the intention of building a global network of volunteer operators. Whilst the project has produced a number of useful analyses of filtering events, the project is still in its infancy and is subject to a number of significant ethical concerns with respect to the risks to participants in the network [31]; we discuss these in greater detail in §3.8.

Certain types of filtering act less on network traffic in transit, and more on application level or social filtering. King et al. [16] studied manual censorship practices in Chinese long-form blogging, and demonstrated that the Chinese censorship authorities were

chiefly concerned with preventing calls to *collective action* whilst allowing significant levels of government criticism.

2.3 Anomaly Detection

The Tor project maintain a censorship flagging tool, as described by Danezis[6], that assesses the ratio of daily connections on a per-country basis over a seven-day time period. If a country’s ratio of users falls significantly outside of the globally-observed usage trends for Tor based on the fifty largest Tor-using countries, which are presumed not to filter connections, then a day is judged to be of interest.

Principal component subspace analysis was used by Lakhina et al. [17] to identify network-wide anomalies in high-speed networks, where long-term patterns are assumed stable, via data gathered from a restricted set of link-level observation points. This is in contrast to our approach, which identifies per-country anomalies in usage data over a sliding time window in data that is assumed non-stationary.

Several other works have extended or expanded aspects of this approach, notably [27], [32], and [11]. These largely focus, however, on using a small number of network observation points to infer network-wide anomalies, and as such typically begin from relatively low-dimensional data. Our approach specifically focuses on per-observation anomalies across a dataset with several hundred dimensions, representing individual countries, in order to highlight states displaying anomalous behaviour.

In addition, to counter the effects of significant long-term shifts in the underlying data, we also perform analyses over a rolling time window to cancel out large-scale developing patterns. This is discussed in greater detail in §3.2.3.

3 Methodology

In this section we discuss the fundamental techniques underlying our approach, and discuss their application to the dataset we use in the rest of this work.

3.1 Tor

Tor [9] is an approach to anonymous web-browsing that offers realistic compromises between latency, usability, and the strength of the anonymity properties that it provides. The most visible end-user aspect of Tor is the Tor Browser Bundle, which provides a web-browser that both uses the Tor network for transport, and is tailored to reduce identifiability of end users.

Managed by the Tor Project, Tor has developed into a global network of volunteer-run relays that forward traffic on behalf of other users. The network makes use of an *onion routing* approach that build encrypted circuits between relays, preventing most realistic adversaries from linking Tor users to particular streams of traffic exiting the network.

The most significant aspect of the Tor network for the present work is that, by its nature, users' traffic is relayed via third parties. As such, and in addition to its anonymity properties, Tor provides a means to bypass many forms of internet filtering. Censorship circumvention is a core aspect of the Tor Project's goals, and significant ongoing research work[20, 29] is aimed at ensuring that Tor continues to offer means to evade national-level filters.

While the extent and popularity of Tor's use in regions that experience significant levels of filtering, such as China, is open to debate [26], Tor is known to have been blocked actively by a number of states, including China and Iran, that object to its use to bypass local internet restrictions and to act anonymously. Significantly, Tor is also arguably the highest-profile censorship circumvention tool at the international level and has received significant media coverage, making it one of the tools of choice for internet activists.

3.1.1 Tor Metrics Data

Tor's role as a high-profile censorship circumvention network make it a useful indicator of global filtering practices. To support analysis of the tool, the Tor project provide estimated daily per-country usage statistics.

Gathering statistics in the Tor network is inherently a difficult task, as the anonymity properties of

the network preclude the identification of individual users via their connections. Instead, the Tor Project make use of client requests to central *directory authorities* to estimate overall user numbers.

When a Tor client connects to the network, or desires to refresh its view of the network, it connects to one of a small number of directory authorities that store a list of all active Tor relays. These directory servers count the number of requests received each day, and geolocate the requesting IP addresses[19]. The resulting aggregate request statistics are passed to a centralised Tor *metrics portal*, from which data is freely available[23].

It is assumed that each client, on average, will make ten requests per day, and as such the aggregate user statistics are divided by ten to provide a final estimate of usage. This data is averaged across each 24-hour period to provide the average number of concurrently connected Tor clients for that day[24]. Whilst the number of distinct clients per day cannot be estimated with any accuracy, the methodology of the Tor metrics portal provides a sufficiently stable estimate.

From these estimates we obtain a set of 251 time series representing individual countries according to the GeoIP database used by Tor. These time series comprise daily observations ranging from the beginning of September 2011 to the time of writing². From these, we remove those countries whose Tor usage never rises above 500 users, on the basis that with such a small number of users, almost any change in use is sufficient to be considered an anomaly. Our techniques can be applied to all countries in the dataset, however, this runs the risk of creating a significant number of false positives³.

²Earlier data is available, but was gathered using a different methodology and has not yet been analysed with the techniques presented in this paper.

³Whether these anomalies are truly false is debatable. They do represent statistically significant shifts in usage patterns for those countries, which may be of interest for more targeted studies, however with respect to understanding exogenous interventions these anomalies are less useful. One possible solution to this would be, rather than removing such countries entirely, to report their results separately to higher-usage countries.

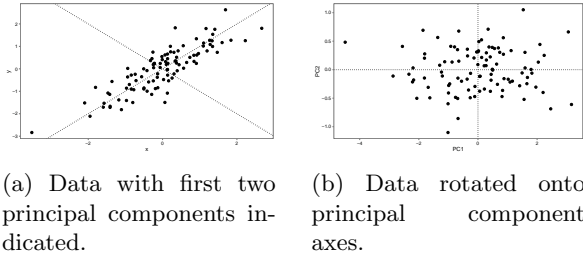


Figure 1: Example principal component transform.

3.2 Principal Component Analysis

Principal component analysis was developed by Pearson[21] as a means to produce tractable low-dimensional approximations of high-dimensional datasets. The original set of variables, which may display correlations, are transformed to a set of linearly uncorrelated variables known as *principal components*.

Principal component analysis transforms a dataset to a new coordinate system in which orthogonal axes, the *principal components*, successively represent the direction of greatest variance in the data. The majority of the variance lies along the first coordinate axis, with the orthogonal coordinate axis that describes the next highest degree of variance being the second principal component. This process continues until the data is fully described by a set of principal components of equal cardinality to the original set of dimensions. Figure 1 demonstrates this transformation on a simple two-dimensional dataset.

When data displays a high degree of correlation between variables then a small number of the most significant principal components may be sufficient to describe the original data to a high degree of accuracy. In many practical scenarios, high dimensional data can be described using only two or three of the most significant principal components.

Before application of principal component analysis, data is typically transformed on a per-dimension basis, to be zero-mean; and scaled to unit variance to ensure that each dimension contributes equally to the result. See [14] for a detailed treatment of principal component analysis and the various choices and compromises to be made when applying the technique.

3.2.1 Application to Tor Metrics Data

Our set of observations can be considered as an $m \times n$ matrix X , in which the n columns correspond to individual countries, and the m rows correspond to date-indexed observations. Each row x_i of X represents a point in n -dimensional space. As noted above, each column is transformed to have a zero mean.

The purpose of a principal components analysis is to form a projection of X into a p -dimensional subspace, where $p \leq n$, such that the sum of the squares of the distances between points and their projection in the subspace are minimized. There are various ways to achieve this, but the most well-known is by maximising the covariance matrix of the projected data.

For our matrix of observations, X , we wish to calculate a set of principal components W , where $|W| = n$. This is achieved by an iterative procedure. The first principal component is defined as the vector that maximises the variance in X :

$$\mathbf{w}_1 = \arg \max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

Subsequent components are those vectors that account for the maximum variance in the residuals between X and the projection of X onto the current set of components. The residuals from the first $k-1$ principal components can be expressed as:

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{X}\mathbf{w}_i\mathbf{w}_i^T$$

3.2.2 Subspace Analysis

As noted above, principal component analysis is a means to project n -dimensional data onto a p -dimensional subspace, which accounts for the majority of the variance in the original dataset. As we are interested in anomalies in data rather than underlying trends, it is possible to consider principal component analysis as dividing the data into two subspaces: a modelled *normal* subspace that accounts for overall trends, and an *anomalous* subspace that is not accounted for by the selected components.

Subspace analysis focuses on the anomalous subspace by inverting the transform using a restricted

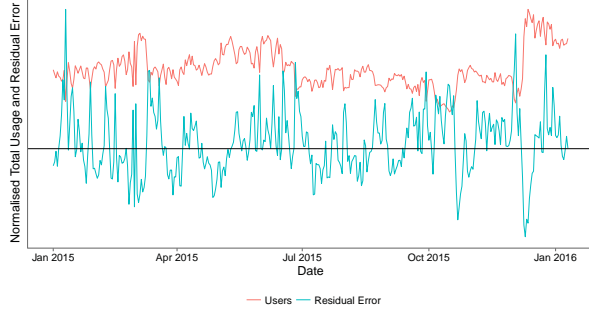


Figure 2: Normalised plot showing daily usage against residual errors for a one-year period in China.

subset of the principal components. This results in an approximation of the data in which the residual errors are the key item of interest, similarly to the original iterative step of the principal component calculation.

In contrast to a reduced-dimensional approximation of the data, we are left with a set of *residuals* that express variances in the data not captured by the chosen set of principal components, and which thus reflect aspects of the data that are not modelled by the more significant components. Intuitively, these residuals reflect behaviour that is not captured by the approximate model. A large-scale residual typically represents behaviour that deviates significantly from previous patterns, and is thus of interest.

We scale these residual errors in proportion to the original data to produce a set of proportional residuals for each of the n countries.

To illustrate, Figure 2 shows the pattern of overall daily usage of Tor China, against the normalised residual prediction error for each day.

3.2.3 Rolling Analysis

As discussed in §1.2, we are interested in evaluating the deviation of individual countries from their predicted patterns over time. In order to highlight developing patterns we therefore do not calculate principal components over the entire time series, which would tend to obscure developing anomalies, but instead perform a rolling principal component analysis

over smaller time windows within the series.

A further justification for this approach is that principal component analysis treats each observation within a time window of X as a point in \mathbb{R}^n , but does not capture the sequential relationship between successive observations. By reducing the time window of observations we focus the analysis, on each successive window, on observations occurring temporally close to each other. This allows the analysis to ignore past and future relative shifts in trends.

We conduct the calculation of residuals only on the final row of each window, corresponding to the most recent observation. Residual anomalies for each day therefore result from the principal components for a prior fixed-length time window.

A principal component analysis requires that there be at least as many data points as there are dimensions in the data, providing a lower-bound on the number of days that must be included in the rolling analysis. At the same time, incorporating too many days into the analysis risks including data that has become irrelevant due to the shift in usage patterns. As a reasonable compromise between short-term sensitivity of results and the dimensionality of the data, all experiments in §4 make use of a rolling 180-day time window.

Our approach makes no assumption that the principal components calculated over each time window are consistent. Whilst the most significant principal components are likely to be similar across time windows, less significant principal components are highly likely to be differing linear combinations of observations. As we are interested in the aggregated principal components in the anomalous subspace, however, this is unlikely to have significant effects on our results. See Ringberg et al.[25] for a discussion of this topic.

3.2.4 Selection of Components

For many applications, a small number of principal components is sufficient to describe a sufficiently large proportion of the variance in a dataset. Lakhina et al. [17] note that multivariate network traffic time series often exhibit low overall dimensionality, and are well described by small numbers of principal

components.

One accepted technique when selecting an appropriate number of components is to calculate the proportion of variance explained by each component, and identify the “elbow” point at which successive principal components add comparatively little extra information about the dataset. This is typically achieved through visual inspection of a scree plot of principal component variance scores.

A second common approach is to make use of Kaiser’s criterion[15] to select only those principal components with eigenvalue greater than 1, representing those components that provide more information than a single average component.

It is worth noting that most approaches to selection of principal components are focused on an appropriate tradeoff between minimising the number of dimensions to be used in further analysis and maximising the explained variance in the data. As the approach taken here is explicitly focused on the residual subspace, however, there is an argument that erring on the side of less principal components, and thus preserving a greater number of residuals, may be of use.

In line with Kaiser’s criteria our experimental results support a value of twelve principal components. As such our tool, and the experiments conducted in the remainder of this work, make use of twelve principal components by default. It is worth noting, however, that this is a potential tuning parameter of the technique, and that a different set of principal components may be more appropriate for other data sources.

Having discussed the various aspects of the approach, we now detail the specific methodology employed in the experiments.

3.3 Applied Subspace Analysis

The previous sections detailed our underlying approach proposed to detecting per-country usage anomalies. We will now discuss how this approach can be applied in practice to the Tor metrics data as the basis for the experimental results shown in §4.

As noted above, our approach differs from most other subspace analyses in that we focus specifically

on anomalies appearing in individual dimensions of the time series, rather than reconstituting system-wide anomalies from a small number of observation points, and makes explicit allowance for shifts in the long term mean of the individual countries in the dataset and their relation to each other.

Our approach allows for global trends in the set of observations to be discounted by considering only anomalies that occur within individual time series outside of global trends. In practice, this allows for internet-wide events, such as global shifts in network usage or large-scale blackouts, not to be identified as anomalies.

This discounting of large-scale events is a significant advantage of our approach over other approaches to anomaly detection that perform univariate time series analyses. By incorporating global patterns of usage, we avoid highlighting anomalous periods that reflect, say, a global botnet or the usage spikes associated with major international events. As we show in §4.1.4, this allows us to focus on those anomalies that are purely local to a given state, or small number of states.

3.4 Filtering Events as Usage Anomalies

It is fundamental to the broader goals of this work that usage anomalies in appropriately selected traffic, and in particular from circumvention tools, can be indicative of the imposition or relaxation of filtering. At the same time, it is clear that other types of event, both technical and sociopolitical, can lead to shifting patterns of usage in these tools. This is, however, a fundamental constraint in analysing this data, and we discuss some implications of this and how to address it in §6. We take the view that any statistically significant shift in a country’s circumvention tool usage is an event of interest to the filtering research community, whether it results from a direct technical intervention, or otherwise.

In our experiments we make use of the Tor metrics data, from which we aim to identify two forms of event: firstly, direct blocking of the Tor network resulting in anomalies in Tor usage; secondly, changing characteristics of Tor usage in response to exoge-

nous factors. The censorship of a major international website, such as YouTube, has the potential to drive a noticeable number of users to Tor, and as such Tor becomes a useful *proxy variable* [28] for a broader class of filtering behaviour.

The direct effects on usage data from filtering events may be expected to fall into three major classes:

- Sharp drops in Tor usage, indicative of direct blocking of the network.
- Gradual drops in usage, which maybe indicative of more subtle filtering but may also represent the relaxation of other practices, resulting in a decrease in users that consider the use of Tor a necessity.
- Sharp spikes in usage, indicative of the blocking of key resources such as major websites that drive a large number of users to the Tor network.

When judging a particular effect as anomalous, the nature of the specific class of anomaly should be considered. For more extreme events, such as the blocking of Tor directly, or the resulting rise in Tor usage due to the blocking of major internet services, large residuals should be expected. In contrast, the subtler effects of media reporting and general internet policy changes may result in subtler and longer-term anomalies. We leave consideration of this class of anomalies to future work.

3.5 Classifying Anomalies through Residuals

The residual errors calculated during the subspace analysis account for variance in the dataset that is not expressed by the chosen principle components in the approximate model. This essentially gives us a measure of how much a given country’s Tor usage differs from the majority of world wide Tor usage on any given day. The magnitude and sign of a residual error provides information on the underlying usage and are used to detect anomalous activity.

- A positive residual represents a drop in the actual Tor usage for a given country.

- A negative residual represents an increase in the actual Tor usage for a given country.
- The magnitude of a residual expresses how much a given country varies from the approximated model.

By examining the residuals, rather than the usage, to determine where anomalous usage occurs, we identify periods of activity that vary from the global trends. The magnitude of the residual errors give us a measure on how much the underlying usage varies from the model and thus is used to identify anomalous activity. Furthermore, we can choose to look at separate classes of anomaly by focusing on the sign of the residuals.

3.5.1 Anomalous patterns in flat usage trends

An advantage of identifying anomalies from residual errors, rather than raw usage numbers directly, is that it incorporates the expected trend of the country. This allows periods be identified as anomalous when they show no rise in usage where a rise would have been expected; even a small rise where a large rise is expected can signify an effective decrease in usage of a circumvention tool.

This capacity to identify anomalies in seemingly typical usage is an important and unusual aspect of our technique, taking advantage of the relative patterns of usage between countries.

An assumption of our approach is that while overall trends in usage of Tor evolve, relative patterns of Tor between countries are largely consistent over shorter time periods; this view is supported by our experimental results. As a result, we avoid manual selection of large countries as a baseline for Tor usage, and instead employ principal component analysis to identify trends directly.

In the remainder of this work, we evaluate our approach against synthetically injected anomalies in the data to analyse the effectiveness of our detection methods as the magnitude and severity of the anomalies vary. We also conduct a series of analyses of the Tor metrics data to identify anomalous countries and specific periods of anomalous behaviour. Finally, we

compare our detection mechanism against the small number of verified reported blocking events against the Tor network.

3.6 Expected Error and Anomalous Threshold

A key element in the approach presented in this work is to determine an appropriate threshold for events to be considered anomalous. The size of this threshold value is inherently linked to the expected error in the technique. We here discuss and justify our approach to calculating this threshold, making use of robust statistics[12] to minimise false detection rates.

Our technique relies on predicting an individual country’s usage for a given day through a linear combination of the usage of other countries for that day, as defined by the principal components. The discrepancy between the prediction and the observed value defines the residual error. Sufficiently large errors are considered anomalous.

3.6.1 Fixed Threshold

A naïve anomalous threshold can be defined simply as a proportion of the usage for that day. If the predicted value falls outside of a given percentage of the true value, a period can be marked as anomalous.

This approach is, however, problematic for a number of reasons. Perhaps the most critical for our application is that different countries in the dataset may be predicted more or less accurately on average than others. As such, countries that are typically predicted poorly by the model, for some reason, will produce a high proportion of anomalous periods.

3.6.2 Dynamic Threshold

To avoid the problems with a fixed threshold, we calculate a threshold based on the characteristics of each country. By tracking the expected residual value over time for each country, with the assumption that anomalies are relatively rare events, an expected anomalous threshold can be determined based on the likelihood of a given observed error for that country.

In data where errors are assumed to be normally distributed, the appropriate means to calculate such an error is through the use of a rolling mean average error and standard deviation about the mean. This allows for the typical variance of the data to be accounted for in the analysis, and anomalies defined as those that fall outside of a number of standard deviations from the mean.

Unfortunately, the mean and standard deviations are not *robust* against outliers in the dataset as they rely on the assumption that errors are gaussian. We instead make use of the *median absolute deviation about the median* (MAD) to define the expected error in normal usage [18].

In contrast to the mean and standard deviation, the median is robust against outliers in the dataset; a small number of extreme events do not significantly alter its value. Similarly, by taking the median of the absolute deviations about the median as a measure of the statistical dispersion in the dataset, we avoid these anomalies from overly affecting the remaining data points.

This approach allows the threshold for anomalies to be expressed in terms of a number of median absolute deviations, similarly to typically used standard deviations about the mean. As a default, we consider events as anomalous if they fall outside of 2.5 median absolute deviations⁴ from the rolling median value. See [18] for a discussion of the robustness of the median and MAD against outliers, and a justification of a 2.5 median absolute deviation threshold.

3.7 Ranking of Countries

As anomalies are judged according to the size of the residual error from the principal component analysis, this provides a convenient metric by which to rank countries according to the level of anomalous behaviour that they exhibit in a given time period.

Following the reasoning established in the previous section, we make use of the size of the median absolute deviation about the median as our metric to rank countries. The output of the ranking for the entire

⁴Corresponding to roughly one expected false positive every 80 days. See §4.1.2 for an experimental analysis of false positives in our approach.

observed time period is shown in Figure 5, and similar outputs can usefully be produced for arbitrary dates and time periods.

3.8 Ethics

Conducting research into network filtering presents a number of ethical issues [31]. The most significant of these is that approaches to investigating network filtering may require direct access to filtered networks. In practice this often involves the participation of in-country experts to conduct local network tests.

Due to the uncertain legal, or quasi-legal, status of violating or investigating state-level network filters, it is generally impossible to quantify the risks to research participants in carrying out network tests. The classic models of informed consent used in many other fields of research can be difficult to apply for a number of reasons. Firstly, approaches to broad-scale network testing preclude intensive training of research participants due to the time and resource constraints, and providing a disclaimer warning of possible risks is not regarded by most ethics bodies as an appropriate level of informed consent.

Secondly, the possible implications of a user’s device being identified as conducting filtering tests are potentially severe, and hard to quantify. Whilst an attempt to perform a DNS request for a suspected blocked social networking site may be considered relatively innocuous by the remote researcher, it may be far more significant to the censor. More seriously, testing for blocking of socially unacceptable or illegal content, such as hate speech or images of sexual abuse, carry more obvious risks to the participant.

As such, where possible, research into network filtering should make use of passive measurements and existing available data sources. The work in this paper is a deliberate attempt to maximise the effectiveness of such a passive approach.

4 Validation

In this section, we judge the efficacy of our method in terms of its ability to detect anomalies, in a variety of circumstances, as well as its false classification rate.

4.1 Error

We discuss here the various possible errors with our approach, and means to analyse them.

4.1.1 Ground Truth

It is extremely difficult to obtain ground truth for internet filtering events, nor are there publicly-available comprehensive lists of filtering. Filtering is, by and large, an opaque process that is rarely announced. Even when states do choose to filter connections openly, the details of that filtering are rarely made public. Existing data therefore relies on word-of-mouth reporting, supported by manual investigation, which has resulted in a dearth of reliable or comprehensive data concerning global filtering.

This poses a problem when attempting to calculate error in our approach, as we have no objective data against which to correlate identified events. Whilst we can attempt to correlate flagged anomalies with events that have been reported and discussed elsewhere, we cannot claim with certainty that we have detected all anomalies.

In the the following sections we attempt to address this lack both by injecting synthetic anomalies into the data, as well as comparing anomalies detected by our method against an existing lists of filtering events.

To evaluate our method, however, we must examine both false positive and false negative rates. A false positive for this approach is be a period where there is no anomalous activity, but anomalous activity is detected; a false negative is a period in which there is anomalous activity but is is not detected.

4.1.2 False Positives

As with any unsupervised machine learning technique the issue of false positives is difficult to address; if our approach detects a period of anomalous behaviour, we have no objective means to ascertain that that period was not anomalous.

Our approach analyses real world data, and any activity that is flagged as anomalous is flagged as such because that time period within the data itself is anomalous. As there is no complete list of all

anomalous events it is impossible for our approach to judge any observed anomaly as a false positive.

We similarly cannot make claims about the causes of detected anomalies. This work considers the correlation between filtering and fluctuations in the number of users of censorship circumvention tools, however there are other events that can create such fluctuations, and these would equally lead to an period being flagged.

A natural example of this could be a nation's internet failing, resulting in a total lack of connections from that nation. Our approach would detect and flag this as an anomaly, however this is clearly not a censorship event⁵.

4.1.3 False Negatives

Some tampering may be too small or subtle to be detected with our method. While many of the censorship events that make the news a wide sweeping filtering of specific sites or content, some will be minor networking or approach changes that are not immediately obvious. It is important to assess the strength of our approach by determining how large an anomaly must be before it is detected.

One method of doing this is by generating anomalies and injecting them into a period that is otherwise anomaly free. This allows us to change the scale and properties of the anomalies that we inject, to determine how large an anomaly has to be before it is detected. We take this approach in §4.2.

An alternative test for false negatives is to compare the results from our method with an external list of censorship events. This allows us to test whether periods exist in which we did not detect anomalous behaviour during a period where external sources believe an event occurred. We take this approach in §4.3.

⁵If, however, this was not a failure but a physical shutdown on behalf of a government, we would judge this as censorship. This highlights the impossibility of detecting filtering or censorship purely from network data.

4.1.4 Discounting of Global Anomalies

A significant aspect of our approach is its ability to discount global effects in the time series in favour of small local effects. This is most notable in the Tor usage statistics in the period beginning in mid-August 2013, at which point a large-scale botnet known as Sefnit began to use the Tor network for its command-and-control infrastructure [8].

In this period, as shown in Figure 5, global Tor usage experienced a sharp increase of almost an order of magnitude during this period. Despite this extreme alteration in global usage patterns, however, the botnet and its subsequent decline are not, for the majority of countries, highlighted as anomalous periods.

4.2 Detection of Injected Events

In this section we inject anomalous events into the data, and observe whether these events are detected and flagged as anomalous. This allows us to determine how great the effect of an anomalous event must be before it is detected. It also allows us to test for false negatives, by determining whether either periods of negativity still display as negative after we have injected an anomalous event into them, or whether the injected event has been correctly identified. False negatives are periods where there is an anomalous event that has not been detected.

We choose to inject our anomalies into for an 18 month period between 21st August 2013 and 21st February 2014 into the timeline of Belgium. We choose this period and country for three reasons:

- There are a relatively stable amount of users relative to the total, yet there is still a large amount of variation present thanks to the introduction of the Sefnit botnet spike (see section 4.1.4). This allows our injected anomalies to be tested over periods of both stability and flux.
- Our method does not detect any anomalous activity over this period. This is important to ensure that when we detect anomalous activity, that detection is due to the injected anomaly

and not due to an anomaly that already existed within the data.

- There has been no evidence or reports that we are aware of that give any indication to a genuine anomaly in Belgium during this time period.

For these experiments we consider our method to have correctly identified an injected anomaly if over 50% of the days where the injected anomaly is occurring is classified as anomalous. This ensures that we only mark a period as anomalous if we detect a majority of its days as outside of typical patterns. This is a relatively conservative approach that could arguably be relaxed in real-world usage.

We generate our fake anomalies based on properties observed in the real data. The strength of the injected anomalies is based on the average daily users for that country, and magnified upwards or downwards gradually to create the anomaly. The anomalies we inject either reduce or increase the number of users by between 0% and 100%, in either a positive or negative direction. They apply their magnifier gradually, ranging over periods of 7 and 49 days. The experiment was ran 1000 times over a range of different magnitudes, including injected anomalies that both positively and negatively impacted the amount of users.

Anomalies with a greater or more sudden effect are easier to detect, while anomalies that are small or gradual are more difficult. To counter this, the generated anomalies are less extreme than many anomalies identified in the real world.

4.2.1 Results

The results of these experiments are shown in Figure 3, showing the changing success rate of detection as the strength of the anomaly increases. Over 50% of anomalies are detected when the anomaly has a positive magnitude of 20% of the total; this rises to a 90% detection rate as the magnitude of the positive anomaly approaches 50% of the unaltered value.

Anomalies with a negative magnitude are detected with a slightly smaller change in the absolute numbers. Over 90% of cases are detected when anomalies

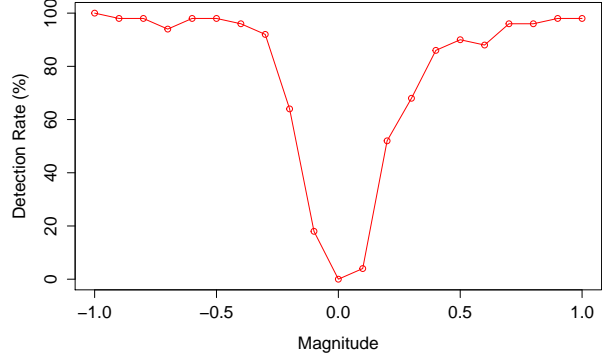


Figure 3: Changing rate of detection as the magnitude of injected anomalies increased.

reduce the number of users by 30%. At a 50% reduction, 98% of anomalies are detected⁶.

4.3 Detection of Known Events

Having calculated anomalous statistics over the historical Tor data set, we now aim to validate our approach by comparing detected anomalies against countries and periods in which internet restrictions are known to have been applied, or in which significant events were occurring that may have influenced usage of circumvention tools.

As noted above, there is a significant lack of ground truth against which to verify our approach. However, it is desirable to validate against an known filtering events whenever possible. To avoid arbitrary selection of favourable cases, however, we restrict our analysis to an externally generated list of reported events.

For this purpose we use [1], a list of reported and verified filtering events against the Tor network dating from 2008 to 2015. This list includes a brief description of each reported event, the dates when the

⁶It is worth noting that a large number of reported events demonstrate increases or decreases in Tor usage that are many orders of magnitude larger than the synthetic events we inject in this analysis.

event was first reported, and how the method was resolved.

This list is sadly brief, reflecting the lack of data available concerning this topic. As mentioned above, one goal of the technique presented here is to provide a baseline of reliable indicators to allow for anomalies to be identified and investigated more thoroughly.

The Tor Project’s metrics data does not cover the full range of events listed in [1]. For those events that do fall within the available data, we analyse here whether these would be detected by our approach.

We consider to have successfully detected a reported anomaly if it is flagged coinciding with the reporting time period. It could be argued that this range should be extended, as a reported anomaly is likely to have started before the report, however we restrict ourselves to this narrow criterion.

As shown in Figure 4, there are only eight reported events that coincide with the published data. Of these, our tool successfully detected five events, and was unsuccessful in detecting a sixth.

For the remaining two events, blocking was carried out against a Tor’s bridge nodes, which are aimed at users who cannot access the normal Tor relays. Data relating to these events falls into a separate dataset not analysed in our results. Bridge node data is noticeably more sparse, due to the much smaller number of users, with only 40 countries in the dataset meeting our usual minimal criteria for numbers of users in a country. We therefore fail to detect anomalous usage in these periods, suggesting that interference with bridge nodes did not significantly affect typical Tor usage at the time.

4.4 Most Anomalous Countries

Figure 5 illustrates the ten most anomalous countries according to their median absolute deviation from the median. Shaded regions denote periods of anomalous usage, according to our tool.

In order, the ten highest-scoring countries over the period from September 2011 until August 2016, and the median absolute deviation of their residuals errors, are shown in Figure 6. While this list includes several of the most well-known filtering states, there are several members of the list that have not received

Country	MAD of Residuals
China	0.1075724
Ethiopia	0.09886959
Iran	0.1321582
Bangladesh	0.285012
South Africa	0.03410801
Moldova	0.0637565
Mongolia	0.09190838
Uganda	0.3610505
Kazakhstan	0.1441831
Yemen	0.1146447

Figure 6: Ten most anomalous countries by median absolute deviation of residual score.

significant attention from the filtering community, and are worthy of further investigation.

5 Discussion

As noted several times above, it is a key limitation to the analysis of this approach that there is no objective ground truth against which to validate the detection rates, and particularly the false positive rate, of our tool. In addition, raw traffic data does not allow for the isolation of filtering events from other forms of intervention, such as network outages or even social and political unrest. In §6, we discuss the implications of this and potential avenues of mitigation.

Despite this, the validation and results of §4 do strongly suggest that the analysis here is practically useful. Whilst we cannot objectively identify censorship detection from this approach we do, as expressed in §1, claim that the highlighted anomalies detected by our approach are strong indicators of regions of likely interest to the internet filtering research and activist communities.

More directly, the experimental validation in the previous section demonstrates that the approach presented here does detect, with reasonable success, a large number of anomalies with varying magnitudes and durations. Again, whilst this cannot conclusively claim identification of internet filtering, we suggest that it provides a valuable practical tool to highlight

Date	Country	Description of Event	Detected
2014-03-28	Turkey	Tor website blocked.	✓
2012-12-16	Syria	DPI on TLS renegotiation.	✓
2013-01-30	Japan	Bridge blocked.	- ¹
2012-10-18	Iran	DPI on TLS for client key exchange.	×
2013-03-09	Iran	SSL handshake filtered.	✓
2014-07-29	Iran	Block directory authorities.	✓
2013-03-26	China	Active probing obfs2 bridges.	✓
2015-02-01	China	Obfs4 default bridges blocked.	-

¹ See §4.3 for a discussion of these bridge node anomalies.

Figure 4: Detection of reported Tor blocking events.

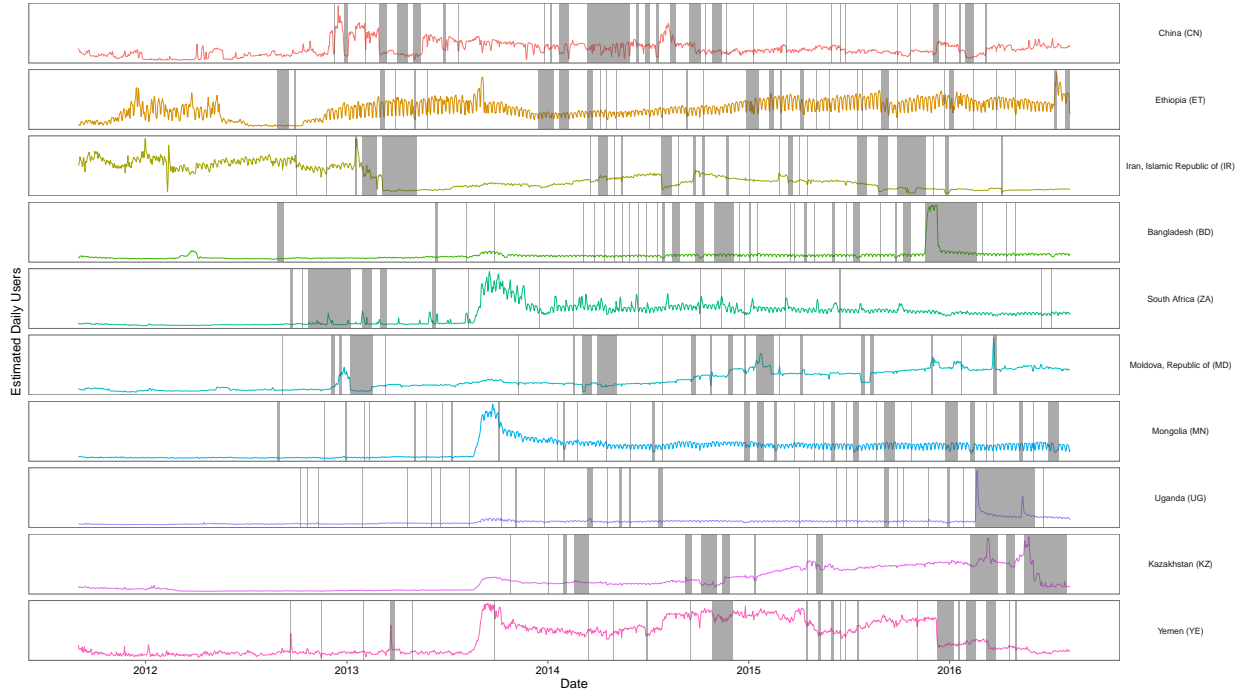


Figure 5: Ten most anomalous countries according to median absolute deviation of residuals over observed time period.

regions and periods of interest as the starting point for more targeted investigations.

6 Future Work

We have demonstrated that principal component analysis can detect various forms of censorship-related events. It would be useful to classify these types of event more explicitly, and to determine appropriate parameters for detecting different classes of filtering-related event.

Whilst the work presented here has focused on the application of our technique to Tor metrics data, the method is more generally applicable. Applying the techniques presented here to other data sources is the most obvious direct extension to this work, and we have made preliminary analyses based on Tor bridge data, data from the Measurement Lab[10], as well as evaluating data from the OONI Project[22] for its applicability in detecting filtering. Other data sources, such as social media, are also likely candidates for analysis.

Our main planned extension to the current work is to develop means to combine multiple data sources into the model, both to strengthen the power of the predictions and to allow events such as network outages to be discounted in the analysis so as to focus more fully on internet filtering events.

A key line of enquiry is to adapt this approach to more sophisticated techniques than principal component analysis. Several similar but improved techniques allow for more effective consideration of the time domain in the analysed data, and the incorporation of multiple data sources per country, allowing for discounting of extraneous factors such as network outages.

7 Conclusions

In this work we have presented an approach based on principal component analysis to detect anomalous periods in per-country usage statistics from globally-gathered time series. We demonstrate that application of this approach to statistics gathered by the

Tor Project’s metrics portal provides a means to indicate potential filtering and censorship-related events at the global level.

To our knowledge, this work provides the first generally applicable tool for detecting a broad class of internet filtering events on a global scale, without the need to focus on individual countries and dynamically adapting to changing patterns of usage. Countries exhibiting anomalous behaviour are automatically identified, and can be subjected to further, more targeted, investigation.

We have validated our approach by correlating detected anomalous periods with known events, and have also demonstrated the effectiveness of the technique by injecting artificial anomalies into an otherwise normal time series and demonstrating the sensitivity of our approach to these anomalies.

Beyond the technique itself, the analyses presented in this work have identified several states that are known to engage in active filtering, but have also highlighted patterns of anomalous behaviour in several states that have not received significant attention from the internet censorship research community. Conducting more detailed investigations of these countries is a promising focus for future research.

In addition to the underlying technique and tool developed to detect anomalous periods of behaviour, we have suggested, and provided initial evidence, that the use of the Tor metrics data, amongst other sources, is of use not only as an indicator of its own usage patterns, but as a practical proxy variable for a much wider class of political and social events. This presents significant potential for researchers, policy makers, and activists investigating global freedom of expression.

A Further Outputs

A.1 Demonstrated Residuals

Figure 7 illustrates the same dataset as Figure 5, but showing a normalised plot of the residual error rather than simply highlighted anomalous periods.

A.2 Recent Anomalies

Figure 8 demonstrates the most highly-ranked countries according to median absolute deviation of residuals, in the most recent 30 days at time of writing.

References

- [1] S. Afroz and D. Fifield. Timeline of tor censorship. <https://metrics.torproject.org/>. Accessed 18th February, 2016.
- [2] Anonymous. The collateral damage of internet censorship by dns injection. *SIGCOMM Comput. Commun. Rev.*, 42(3):21–27, June 2012.
- [3] Anonymous. Towards a comprehensive picture of the great firewall’s dns censorship. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*, San Diego, CA, Aug. 2014. USENIX Association.
- [4] R. Clayton, S. J. Murdoch, and R. N. M. Watson. Ignoring the great firewall of china. In *Proceedings of the 6th International Conference on Privacy Enhancing Technologies, PET’06*, pages 20–35, Berlin, Heidelberg, 2006. Springer-Verlag.
- [5] J. R. Crandall, D. Zinn, M. Byrd, E. Barr, and R. East. ConceptDoppler: A Weather Tracker for Internet Censorship. *Computer and Communications Security*, Oct. 2007.
- [6] G. Danezis. An anomaly-based censorship-detection system for Tor. Technical report, The Tor Project, 2011.
- [7] R. Deibert. *Access Denied: The Practice and Policy of Global Internet Filtering (Information Revolution and Global Politics Series)*. MIT Press, 1 edition, Dec. 2007.
- [8] R. Dingledine. How to handle millions of new tor clients. <https://blog.torproject.org/blog/how-to-handle-millions-new-tor-clients>. Accessed 18th February, 2016.
- [9] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second-generation onion router. In *IN PROCEEDINGS OF THE 13 TH USENIX SECURITY SYMPOSIUM*, 2004.
- [10] C. Dovrolis, P. K. Gummadi, A. Kuzmanovic, and S. D. Meinrath. Measurement lab: overview and an invitation to the research community. *Computer Communication Review*, 40(3):53–56, 2010.
- [11] L. Huang, X. Nguyen, M. Garofalakis, and J. M. Hellerstein. Communication-efficient online detection of network-wide anomalies. In *IEEE Conference on Computer Communications (INFOCOM)*, pages 134–142. IEEE, 2007.
- [12] P. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2004.
- [13] J. E. Jackson and G. S. Mudholkar. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics*, 21(3):341–349, 1979.
- [14] I. T. Jolliffe. *Principal component analysis*. Springer, New York, 2002.
- [15] H. F. Kaiser. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20:141–151, 1960.
- [16] G. King, J. Pan, and M. E. Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107:1–18, 2013.
- [17] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *Proceedings of ACM SIGCOMM 2004*, pages 219–230, Aug. 2004.



Figure 7: Normalised plot showing daily usage against residual errors for entire time period, for the most highly ranked anomalous countries according to median absolute deviation of residuals.

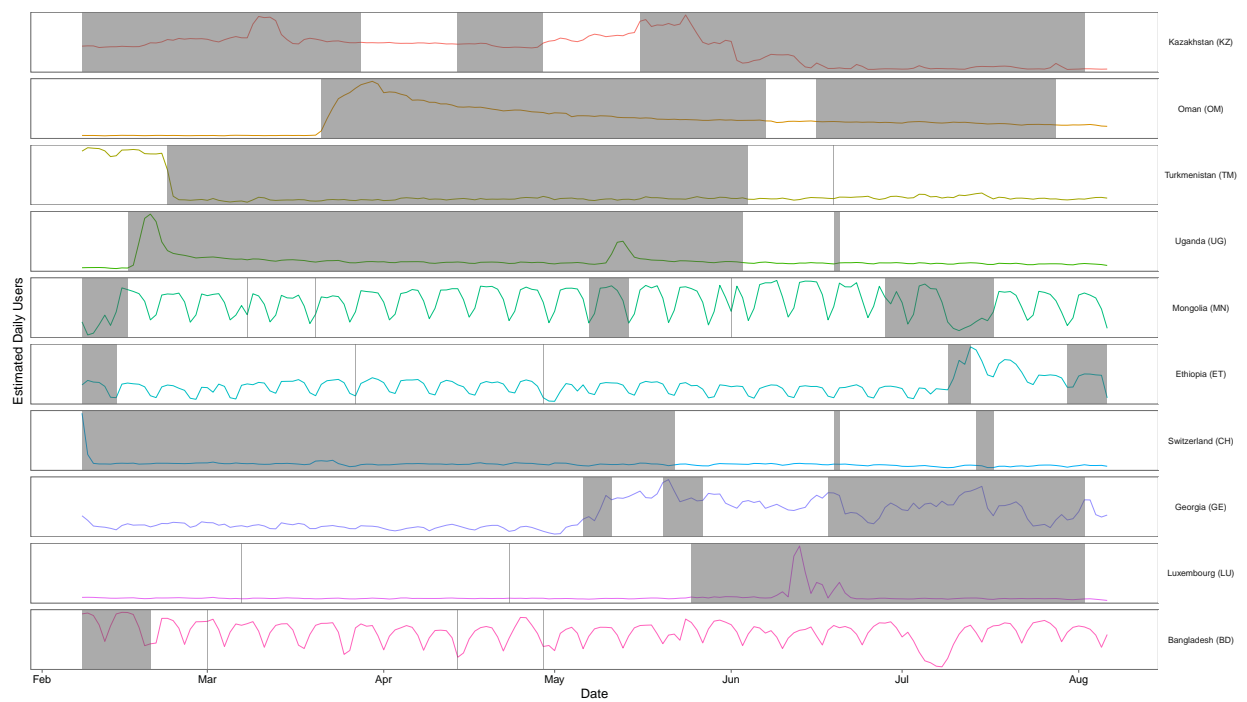


Figure 8: Most anomalous countries according to median absolute deviation, in the 30 days prior to 2016-08-06.

- [18] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764 – 766, 2013.
- [19] MaxMind Inc. MaxMind GeoIP City Database. <http://www.maxmind.com/app/city>. Accessed 14th May 2015.
- [20] H. M. Moghaddam, B. Li, M. Derakhshani, and I. Goldberg. Skypemorph: Protocol obfuscation for Tor bridges. In *Proceedings of the 19th ACM conference on Computer and Communications Security (CCS 2012)*, October 2012.
- [21] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.
- [22] The OONI Project. The open observatory of network interference. <https://ooni.torproject.org/>. Accessed 18th February, 2016.
- [23] The Tor Project. Tor metrics portal. <https://metrics.torproject.org/>. Accessed 18th February, 2016.
- [24] The Tor Project. Tor metrics: Questions and answers about user statistics. <https://gitweb.torproject.org/metrics-web.git/tree/doc/users-q-and-a.txt>. Accessed 18th February, 2016.
- [25] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. In *SIGMETRICS '07: Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pages 109–120, New York, NY, USA, 2007. ACM Press.
- [26] D. Robinson, H. Yu, and A. An. Collateral freedom: A snapshot of chinese users circumventing censorship. Technical report, 2013.
- [27] A. Soule, K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In *In Proceedings of IMC*, 2005.
- [28] G. Upton and I. Cook. *Oxford dictionary of statistics*. Oxford university press Oxford, UK, 2002.
- [29] Z. Weinberg, J. Wang, V. Yegneswaran, L. Briesemeister, S. Cheung, F. Wang, and D. Boneh. StegoTorus: A camouflage proxy for the Tor anonymity system. In *Proceedings of the 19th ACM conference on Computer and Communications Security (CCS 2012)*, October 2012.
- [30] J. Wright. Regional variation in chinese internet filtering. *Information, Communication & Society*, 17(1):121–141, 2014.
- [31] J. Wright, T. de Souza, and I. Brown. Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics. In *Free and Open Communications on the Internet*, San Francisco, CA, USA, 2011. USENIX.
- [32] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan. Network anomography. In *In IMC*, 2005.